(4)

CRM 87-161 / August 1987

RESEARCH MEMORANDUM

PROPERTIES OF SOME BAYESIAN SCORING PROCEDURES FOR COMPUTERIZED ADAPTIVE TESTS

D.R. Divgi

A Division of CNA Hudson Institute

CENTER FOR NAVAL ANALYSES
4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268

Work conducted under contract N00014-87-C-0001

This Research Memorandum represents the best opinion of CNA at the time of issue. It does not necessarily represent the opinion of the Department of the Navy

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a REPORT SECURITY CLASSIFICATION Unclassified	1b RESTRICTIVE MARKINGS
2a SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b DECLASSIFICATION / DOWNGRADING SCHEDULE	
4 PERFORMING ORGANIZATION REPORT NUMBER(S) CRM 87-161	5 MONITORING ORGANIZATION REPORT NUMBER(S)

6a NAME OF PERFORMING ORGANIZATION Center for Naval Analyses	6b OFFICE SYMBOL (If applicable) CNA	7a NAME OF MONITORING ORGANIZATION Commandant of the Marine Corps (Code RDS)
6c ADDRESS (City, State and ZIP Code) 4401 Ford Avenue Alexandria, Virginia 22302-0268		7b ADDRESS (City, State, and ZIP Code) Headquarters, Marine Corps Washington, D.C. 20380
8a NAME OF FUNDING / ORGANIZATION Office of Naval Research	8b OFFICE SYMBOL (If applicable) ONR	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N00014-87-C-0001
8c ADDRESS (City, State, and ZIP Code) 800 North Quincy Street Arlington, Virginia 22217		10 SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO	WORK UNIT ACCESSION NO
65153M	C0031		

11 TITLE (Include Security Classification)

Properties of Some Bayesian Scoring Procedures for Computerized Adaptive Tests

12 PERSONAL AUTHOR(S)
D.R. Divgi

13a TYPE OF REPORT	13b TIME COVERED	14 DATE OF REPORT (Year, Month, Day)	15 PAGE COUNT
Final	FROM TO	August 1987	24

16 SUPPLEMENTARY NOTATION

17 COSATI CODES

FIELD	GROUP	SUB-GROUP
05	08	
12	03	

18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)

ASVAB (Armed Services Vocational Aptitude Battery), Bayes Theorem, CAT (Computerized Adaptive Testing), Computerized simulation, Predictions, Scoring, Statistical analysis, Test methods, Test scores ,

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

The computerized adaptive version of the Armed Services Vocational Aptitude Battery will use a Bayesian procedure for computing test scores. Properties of three common Bayesian procedures are examined in this research memorandum. The results show that the procedures are almost equally reliable and that reliability drops if item parameters change from paper-pencil to computerized administration.

20 DISTRIBUTION / AVAILABILITY OF ABSTRACT □ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT □ DTIC USERS	21 ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED	
22a NAME OF RESPONSIBLE INDIVIDUAL Major Robinson	22b TELEPHONE (Include Area Code) (703) 824-2643	22c OFFICE SYMBOL RDS-40

83 APR edition may be used until exhausted

All other editions are obsolete

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

CNA CENTER FOR NAVAL ANALYSES

A Division of Hudson Institute 4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268 • (703) 824-2000

5 November 1987

MEMORANDUM FOR DISTRIBUTION LIST

Subj: Center for Naval Analyses Research Memorandum 87-161

Encl: (1) CNA Research Memorandum 87-161, "Properties of Some Bayesian Scoring Procedures for Computerized Adaptive Tests," by D. R. Divgi, August 1987

1. Enclosure (1) is forwarded as a matter of possible interest.

2. A computerized adaptive testing (CAT) version of the Armed Services Vocational Aptitude Battery (ASVAB) is being developed for joint-service use by the Navy Personnel Research and Development Center (NPRDC). There are different ways of computing an examinee's CAT score. This Research Memorandum compares three Bayesian scoring procedures - posterior mean, posterior mode, and Owen's approximation - in terms of their reliabilities and of their sensitivity to changes in item parameters from paper-pencil to CAT administration.

William H Sims

William H. Sims
Director, Manpower and Training Program
Marine Corps Operations Analysis Group

Distribution List:
Reverse Page

Subj: Center for Naval Analyses Research Memorandum 87-161

Distribution List

SNDL

A1	ASSTSECNAV MRA
A1	DASN - MANPOWER (2 copies)
A6	HQMC MPR
	Attn: Code M
	Attn: Code MR
	Attn: Code MP
	Attn: Code MA
A6	HQMC TRNG
A6	HQMC RD&S
A6	HQMC RA
A6	HQMC AVN
E3D1	CNR
E3D5	NAVPERSRANDCEN
	Attn: Technical Director (Code 01)
	Attn: Technical Library
	Attn: CAT/ASVAB PMO
	Attn: Director Testing Systems (Code 63)
FF38	USNA
	Attn: Nimitz Library
FF42	NAVPGSCOL
FF44	NAVWARCOL
FJA1	COMNAVMILPERSCOM
FJB1	COMNAVCRUITCOM
FT1	CNET
V12	MCDEC
	Attn: Commanding General
	Attn: Development Center, Plans Division (Code D08) (2 copies)

OPNAV

OP-01
OP-11
OP-12
OP-13

OTHER

Joint Services CAT-ASVAB Working Group (15 copies)
Defense Advisory Committee on Military Personnel Testing (8 copies)

CRM 87-161 / August 1987

PROPERTIES OF SOME BAYESIAN SCORING PROCEDURES FOR COMPUTERIZED ADAPTIVE TESTS

D.R. Divgi

Marine Corps Operations Analysis Group

A Division of CNA *Hudson Institute*

CENTER FOR NAVAL ANALYSES

4401 Ford Avenue • Post Office Box 16268 • Alexandria Virginia 22302-0268

ABSTRACT

The computerized adaptive version of the Armed Services Vocational Aptitude Battery will use a Bayesian procedure for computing test scores. Properties of three common Bayesian procedures are examined in this research memorandum. The results show that the procedures are almost equally reliable and that reliability drops if item parameters change from paper-pencil to computerized administration.

EXECUTIVE SUMMARY

INTRODUCTION

The Department of Defense is developing a computerized adaptive testing (CAT) version of the Armed Services Vocational Aptitude Battery (ASVAB). In CAT, each examinee is characterized by a value of ability, θ; each item is characterized by three parameters which represent discriminating power, difficulty, and the effect of guessing. An experimental version of CAT-ASVAB has been developed and was administered to recruits from all services in a study of CAT validity.

The prior distribution and an examinee's item responses together provide the posterior distribution of that individual's ability. Different scoring procedures (called "estimators" in statistics) can be used for calculating an estimate of the examinee's ability. The purposes of this research memorandum are to distinguish between theoretical and practical criteria for choosing among estimators and to evaluate the psychometric properties of three procedures.

THEORETICAL vs. PRACTICAL CRITERIA

Some researchers have recommended, on theoretical grounds, that the mean of the posterior distribution be used as the ability estimate. Their criterion fcr evaluating an estimator is its mean squared error (MSE)—that is, the average of the squared difference between the true θ and its estimate. In practice, the MSE criterion is irrelevant. The goal of the CAT-ASVAB and of the paper-pencil (PP) ASVAB is not to estimate a parameter θ in a model but to predict future performance. Therefore, CAT-ASVAB will be evaluated in the long run on the basis of its predictive validity. In the short run, it will be judged by the reliabilities of the CAT-ASVAB scores that are used for selection and classification. In particular, CAT subtests should be at least as reliable as their PP counterparts. Reliability and validity of CAT-ASVAB may suffer if an irrelevant criterion is used to select the scoring procedure.

METHODOLOGY

Three Bayesian estimators were evaluated using simulations—that is, computer generation of examinees' abilities and item responses. One estimator was the posterior mean. The second estimator was the mode of the posterior distribution, which is frequently used because it is easier to compute than the mean. The third estimator was Owen's approximation which, despite its simplicity, is known to yield reasonable estimates. MSEs as well as reliabilities were computed for all three estimators.

The simulation imitated the experimental CAT-ASVAB as far as possible, using the same item parameters and item selection algorithm. The standard normal distribution was used as the prior distribution. The true distribution of ability was taken to be normal, with mean and variance equal to estimates based on the recruit sample. Each simulated examinee was administered 10 items in Paragraph Comprehension and 15 in each of the other subtests.

In the experimental CAT-ASVAB project, item parameters were estimated from a PP administration of the item pool and then used in CAT. (The same procedure is being followed in the Accelerated CAT-ASVAB Project.) The implied assumption is that the parameters are not affected by the medium of administration. This assumption is known to be false. Its violation may affect different estimators to different degrees. Therefore a second simulation was performed. The same item parameters as in the first simulation were used for item selection and to calculate all ability estimates. However, while generating examinees' responses, probabilities of correct answers were computed using CAT-based parameter values obtained in an earlier CNA study.

RESULTS

The posterior mean, posterior mode, and Owen's approximation were found to be almost equally reliable (see table I). Results of the second simulation were similar to those of the first in that the three estimators were about equally reliable. Thus, the theoretical superiority of the posterior mean does not translate into a higher reliability than that of the posterior mode. Although Owen's estimator is equally reliable, there is no justification for using an approximation when an estimate based on the correct posterior distribution can be calculated. Thus, the results support using the posterior mode because it is easier to calculate.

Another finding from the second simulation was that changes in item parameters from PP to CAT noticeably reduced reliability. The decreases in reliabilities are presented in table II, where the degree of change in item parameters is indicated by the mean average absolute difference (AAD) between the item characteristic curves in PP and CAT administrations. When a new CAT is being developed, the size of this change is unknown and hence simulations cannot allow for it. As a result, these simulations overestimate CAT reliability.

TABLE I

RELIABILITIES OF SUBTESTS WHEN SCORES ARE COMPUTED USING POSTERIOR MEAN, POSTERIOR MODE, AND OWEN'S APPROXIMATION

	Subtest								
Estimator	GS	AR	WK	PC	AI	SI	MK	MC	EI
Mean	.884	.899	.895	.775	.887	.900	.915	.841	.863
Mode	.884	.898	.894	.773	.886	.899	.914	.839	.862
Owen	.883	.896	.892	.777	.885	.898	.911	.840	.861

TABLE II

SIZE OF CHANGE IN ITEM PARAMETERS AND CONSEQUENT DECREASE IN RELIABILITY

	Subtest								
	GS	AR	WK	PC	AI	SI	MK	MC	EI
Mean AAD	.050	.047	.048	.051	.061	.079	.064	.089	.071
Decrease in reliability	.053	.048	.047	.050	.013	.035	.048	.061	.049

CONCLUSIONS

- Criteria behind technical decisions should be based on the way CAT-ASVAB will be used and evaluated, not on abstract theoretical principles.

- The mode of the posterior ability distribution is a good scoring method for CAT-ASVAB.

- Because of changes in item parameters from PP to CAT administration, reliabilities of CAT-ASVAB subtests will almost certainly be lower than the values obtained in simulations.

TABLE OF CONTENTS

LIST OF TABLES

INTRODUCTION

Within a few years the Department of Defense may begin administering the Armed Services Vocational Aptitude Battery (ASVAB) using computerized adaptive testing (CAT). CAT is based on item response theory (IRT). Each examinee is characterized by a value of ability θ. Each test item is described by an item response curve which specifies how the probability of correctly answering the item increases with ability. The three-parameter logistic model is used in the CAT-ASVAB project. In this model, the probability of a correct answer is given by

$$P(\theta) = c + (1 - c) / [1 + exp\{1.7a(b - \theta)\}] \quad ,$$

where a, b, and c are the discrimination, difficulty, and guessing parameters of the item.

In the experimental CAT-ASVAB [1], which was administered to recruits in all services in the CAT validity study, adaptive testing begins with a highly discriminating item of medium difficulty, selected at random from five such items. The examinee's answer is used to estimate ability, θ. This estimate is used to select the next item to be administered, after which θ is reestimated, and so on. Testing continues until a prespecified number of items has been administered.

A Bayesian procedure is used to update information about θ. One begins with an assumed prior distribution of ability. After the first item the distribution is multiplied by the probability of the examinee's response ($P(\theta)$ for a correct answer, $1 - P(\theta)$ for a wrong one). The product is the posterior distribution of θ (except for a constant factor which is of no consequence). The posterior distribution after the first item is the prior distribution for the second item. When it is multiplied by the probability of the response on the second item, one obtains a new posterior distribution which yields the next estimate of θ. Such sequential updating is continued until a prespecified number of items has been administered.

COMPUTING EXAMINEE'S SCORE

The exact posterior distribution requires extensive calculations. When a microcomputer is used to administer CAT, these calculations may take long enough for the examinee to notice the delay in administering the next item. Therefore Owen's approximation [2] was used in the experimental CAT-ASVAB and will be used in the Accelerated CAT-ASVAB Project [3 (enclosure 3.13, item B3)]. Owen's procedure begins with a normal prior distribution. After each item the correct posterior distribution is replaced by a normal distribution with the same mean and variance, which can be computed using relatively simple formulas. The mean is used as the ability estimate for choosing the next item.

The primary shortcoming of Owen's estimate is that it depends on the order in which items are administered [4; 5 (enclosure 3.3)]. If two persons answer the same items the same way but in different orders, their Owen estimates will not be exactly equal. This is not important as long as the estimate is used only to select the next item. However, the final ability estimate following the last item, after appropriate transformation, becomes the examinee's score on the test. It should be independent of the item order, which is the case with estimates based on the correct posterior distribution.

It is highly improbable, but not impossible, for two persons to be administered the same items in different orders. However, the very possibility is enough to decide the issue. The only justification for using Owen's approximation is that it can be computed much faster than any estimate based on the correct posterior distribution. This is important during item selection because one must not make the examinee wait too long while the interim ability estimate is being computed. However, once a subtest has been completed, there is no urgency about starting the next one. There is enough time to use the correct posterior distribution. Thus, there is no argument in favor of Owen's approximation as the final ability estimate. Hence its dependence on item order, although trivial in its impact on examinees, suffices to rule it out for the final estimate.

The two popular Bayesian estimators are the mean and the mode of the posterior distribution of ability. However, they cannot be reported to test users. The ASVAB has its own score scale based on Form 8a, and it must be used to report CAT-ASVAB scores. Therefore, each CAT-ASVAB subtest score will be equated to an 8a score. As the first step in this equating, the ability estimate will be converted into the expected number right score on Form 8a [3 (enclosure 3.13, item E2.1)]. The objective of this paper is to compare the mode and the mean as estimators, in both θ and number-correct metrics. Owen's estimate, in the θ metric only, is included because it was used as the final score in the experimental CAT-ASVAB system [1 (p. 5)].

Before one can evaluate and compare estimators (i.e., procedures for scoring the test), one must choose a criterion. A criterion based on statistical decision theory differs from one based on practical and psychometric considerations. The distinction is explained in detail below, because if the choice of a scoring procedure is based on an irrelevant criterion, the usefulness of CAT-ASVAB may suffer.

THEORETICAL vs. PRACTICAL CRITERIA

Let $\hat{\theta}$ be an estimator of θ. $E(\hat{\theta} | \theta)$ represents its expected (i.e., mean) value in a subpopulation of examinees, all of whom have ability θ . In general, this mean does not equal θ. The difference is bias B, which depends on θ. Thus, one can write

$$\hat{\theta} = \theta + B + e \ , \qquad (1)$$

where e is random error. Mean e is zero at each value of θ but its variance and the shape of its distribution may depend on θ. The error of estimation is $(B + e)$ and the mean squared error over the entire population of examinees is

$$MSE(\hat{\theta}) = E(B^2) + \text{Var}(e) \ .$$

Bock and Mislevy [6] and Sympson [7] have argued that the mean of the posterior distribution should be used for estimating ability in CAT, because it is the estimator with the smallest MSE. The argument is invalid for two reasons. First, it is based on three assumptions: (1) the three-parameter model is correct; (2) the item parameters are known exactly; and (3) the prior distribution equals the true distribution of ability in the population. In practice, all three assumptions are false to some extent.

The second reason, which is more important than the first one, is that the MSE criterion is irrelevant to CAT-ASVAB. The goal of CAT-ASVAB is to predict future performance, not to estimate a parameter in a theoretical model. Therefore CAT-ASVAB will be evaluated in the long run on the basis of its predictive validity. In the short run it will be judged by the reliabilities of its scores. Hence, in this study, comparisons of estimators are based on concepts of classical test theory.

In classical test theory, the score X on a test is an estimate of the examinee's true score T which, by definition, is the mean $E(X | \theta)$ one would obtain if one could test the examinee repeatedly. (Equivalently, it is the mean over the subpopulation of examinees with the same θ.) Therefore the examinee's true score depends on the procedure used to score the test. The difference between X and T is the error of measurement:

$$X = T + e \qquad (2)$$

with $E(e|T) = 0$. The reliability of X is

$$R = \text{Var}\,(T)/[\text{Var}\,(T) + \text{Var}\,(e)] = 1 - \text{Var}\,(e)/[\text{Var}\,(T) + \text{Var}\,(e)]$$

When the score X is an estimate $\hat{\theta}$ of θ,

$$T = \theta + B \ .$$

The random error e in equation 1 is the same as the error of measurement e in equation 2.

Thus, the role of bias in evaluation of a scoring procedure depends on what is being estimated. If X is considered an estimate of the model parameter θ, B is a part of estimation error and minimum MSE is a legitimate criterion for choosing among estimators (as in [6, 7]). If X is considered an estimate of T, B is a part of the examinee's true score and the criterion is maximum reliability. The latter, not the former, is the role of the test score in mental measurement.

The two criteria, MSE and reliability, may yield different conclusions. This can be seen with a trivial example. Suppose $\hat{\theta}$ is the posterior mean. Then 10 times $\hat{\theta}$ has a much larger MSE. However, since Var(T) and Var(e) both are multiplied by 100, the reliability of $10\,\hat{\theta}$ is the same that of $\hat{\theta}$. Except in the case of very simple models, reliability of an estimator cannot be calculated theoretically. Simulated or real data are needed.

SIMULATION

The simulation attempted to imitate the experimental CAT-ASVAB as far as possible. Item parameter estimates for the experimental CAT-ASVAB item pool were used as the true item parameters. The information table contained 37 equally spaced ability values from –2.25 to 2.25. The "54321 strategy" was used to randomize the choice of the first four items in each subtest [1 (p. A12 and Supplement, p. 91)].

For each subtest, 2,000 abilities were sampled from a normal distribution; mean and standard deviation of the normal were the estimates obtained for the recruit sample that took the experimental CAT-ASVAB [8]. However, in keeping with the experimental system, the standard normal distribution (abbreviated as $N(0,1)$) was used as the prior distribution in all calculations.

The adaptive subtests in CAT-ASVAB are General Science (GS), Arithmetic Reasoning (AR), Word Knowledge (WK), Paragraph Comprehension (PC), Auto Information (AI), Shop Information (SI), Mathematics Knowledge (MK), Mechanical Comprehension (MC), and Electronics Information (EI). Each examinee was administered

10 items in PC and 15 items in all the other subtests. Posterior mean and mode were calculated at the end of each subtest.

As mentioned earlier, CAT-ASVAB ability estimates will be converted into expected number-correct scores on ASVAB Form 8a. As this transformation was not available for all subtests, it was imitated as follows. For each subtest, parameters of all items in the pool were averaged to obtain mean values $\bar{a}$, $\bar{b}$, and $\bar{c}$. These were used to transform the posterior mean to the percent-correct metric:

$$P(\text{mean}) = 100\bar{c} + 100(1-\bar{c})/[1+\exp\{1.7\bar{a}\ (\bar{b}-\text{mean})\}]\ .$$

P(*mode*) was calculated similarly. Two other scores were obtained by transforming the posterior distribution to the percent-correct metric first, and then computing its mean and mode. These will be denoted by *mean*(*P*) and *mode*(*P*). The parameter (*not* the true score) being estimated by all four of these scores is $P(\theta)$, the percent-correct transform of θ.

Mean squared error and correlation with the corresponding parameter were computed for each score. For each estimator, true score as a function of the relevant parameter (θ or $P(\theta)$) was estimated by cubic regression. Reliability was estimated as the multiple R-square of this fit and then used to calculate variances of true scores and measurement errors.

RESULTS

Table 1 presents results for estimators in the θ metric. Posterior mean does *not* have smaller MSE than the mode. This happens because the $N(0,1)$ prior distribution differs from the marginal distribution, that is, the distribution of ability in the population. Thus, the theoretical superiority of the posterior mean applies only when the test is administered to one specific population.

When the classical criterion of reliability is used, the mean and the mode are found to be almost equally good. Surprisingly, in spite of the drastic approximations involved, Owen's estimate is practically as reliable as those based on the exact posterior distribution.

The squared correlation between θ and the estimator is often used as the measure of reliability. Table 1 shows that this underestimates reliability by a small amount. The difference occurs because bias is a nonlinear function of θ.

TABLE 1

PROPERTIES OF ESTIMATORS IN θ METRIC WITH $N(0,1)$ PRIOR DISTRIBUTION, MODEL ASSUMPTIONS SATISFIED[a]

Sub-test	Mean squared error of estimation			Variance of measurement error			Variance of true scores			Squared correlation with θ			Reliability		
	Mean	Mode	Owen	Mean	Mode	Owen	Mean	Mode	Owen	Mean	Mode	Owen	Mean	Mode	Owen
GS	4.7	4.5	5.0	4.5	4.4	4.7	34.0	33.3	35.0	.882	.883	.881	.882	.883	.881
AR	4.4	4.1	5.1	4.3	4.0	4.8	37.4	34.8	39.2	.896	.896	.890	.897	.896	.892
WK	3.7	3.6	4.2	3.7	3.6	4.0	30.7	29.6	31.8	.893	.893	.888	.893	.893	.888
PC	10.3	9.0	11.1	9.8	8.6	10.3	32.3	28.2	34.8	.764	.763	.766	.767	.766	.772
AI	6.6	6.5	7.2	6.2	6.0	6.6	47.3	45.1	48.5	.880	.879	.873	.884	.883	.881
SI	7.9	7.9	8.5	7.2	7.0	7.4	64.2	62.0	64.0	.896	.895	.889	.899	.898	.897
MK	4.1	3.8	4.6	3.9	3.7	4.2	39.6	37.4	40.7	.909	.908	.904	.910	.910	.906
MC	7.8	7.0	8.5	7.3	6.5	7.6	35.2	31.0	36.6	.819	.818	.815	.828	.827	.828
EI	8.5	8.5	8.8	7.8	7.7	7.9	48.2	47.6	47.5	.860	.859	.854	.861	.861	.858

[a] Mean squared errors and variances have been multiplied by 100.

To see if the mean has superior reliability when the prior distribution is correct, that is, equals the marginal distribution, the three ability estimates were recomputed for each examinee using the correct prior distribution. The results are shown in table 2. The variance of measurement error falls when the correct prior distribution is used, but so does the true-score variance, with the result that reliability increases only slightly. The posterior mean does have smaller MSE than the mode, but its reliability is not superior by more than .002. Thus, like table 1, table 2 shows that the three estimators are about equally reliable. Therefore, to keep the simulations realistic, all further calculations use the $N(0,1)$ prior distribution as in the experimental CAT-ASVAB.

Table 3 presents results for scores in the percent-correct metric. They show the same patterns as in table 1. *Mean*(P), which is the posterior mean computed after transformation to the percent-correct metric, is slightly more reliable than the other three. Squared correlations with the parameter $P(\theta)$ are not presented because they differed very little from reliabilities.

SIMULATION WITH MEDIUM-OF-ADMINISTRATION EFFECT

The results shown in tables 1 to 3 are based on simulations in which assumptions of item response theory were satisfied. Different results may be obtained when assumptions are violated. In an operational CAT project, item parameters are estimated from PP administration of the item pool and then used in CAT. This assumes that the parameters are not affected by the medium of administration. This assumption is known to be false. Using data from the recruit sample to which the experimental CAT-ASVAB was administered, Divgi and Stoloff [8] found that observed $P(\theta)$ differed substantially from those calculated from the PP-based item parameters. Therefore, a second simulation was performed in which parameters were changed from PP to CAT for those items that had been answered by at least 1,000 recruits and hence had CAT-based parameters estimated by Divgi [9].

As in operational CAT, PP-based item parameters (that is, those used in the first simulation) were used for item selection and to calculate all ability estimates. However, while generating examinees' responses, probabilities of correct answers were calculated using parameter values based on CAT administration [9].

Tables 4 and 5 contain results in θ and percent-correct metrics, which are similar to those in tables 1 to 3 as far as comparisons among estimators are concerned. It is the comparison between tables that is interesting. The effect of the medium of administration

TABLE 2

PROPERTIES OF ESTIMATORS IN θ METRIC WITH PRIOR AND MARGINAL DISTRIBUTIONS EQUAL, MODEL ASSUMPTIONS SATISFIED[a]

Sub-test	Mean squared error of estimation			Variance of measurement error			Variance of true scores			Squared correlation with θ			Reliability		
	Mean	Mode	Owen	Mean	Mode	Owen	Mean	Mode	Owen	Mean	Mode	Owen	Mean	Mode	Owen
GS	4.2	4.2	4.2	3.6	3.7	3.6	27.6	27.8	27.4	.883	.884	.882	.884	.884	.883
AR	4.0	4.0	4.2	3.3	3.2	3.4	29.8	28.7	29.3	.896	.895	.890	.899	.898	.896
WK	3.3	3.4	3.4	3.0	3.0	3.0	25.3	25.1	24.9	.895	.893	.890	.895	.894	.892
PC	7.9	8.0	7.9	5.9	5.7	5.8	20.5	19.4	20.2	.768	.766	.767	.775	.773	.777
AI	6.3	6.5	6.6	5.3	5.2	5.3	41.5	40.5	40.9	.881	.879	.874	.887	.886	.885
SI	7.8	8.0	8.4	6.6	6.5	6.6	59.3	58.0	58.1	.895	.895	.888	.900	.899	.898
MK	3.5	3.6	3.8	3.0	2.9	3.1	32.1	31.3	31.6	.910	.909	.905	.915	.914	.911
MC	6.3	6.5	6.5	4.7	4.5	4.7	25.1	23.4	24.8	.824	.823	.819	.841	.839	.840
EI	8.3	8.5	8.6	6.9	7.0	6.9	43.6	43.7	42.4	.860	.860	.855	.863	.862	.861

[a] Mean squared errors and variances have been multiplied by 100.

TABLE 3

PROPERTIES OF ESTIMATORS IN PERCENT-CORRECT METRIC WITH *N*(0,1) PRIOR DISTRIBUTION, MODEL ASSUMPTIONS SATISFIED

	Mean squared error of estimate				Variance of measurement error				Reliability			
Subtest	P(mean)	P(mode)	Mean(P)	Mode(P)	P(mean)	P(mode)	Mean(P)	Mode(P)	P(mean)	P(mode)	Mean(P)	Mode(P)
GS	45.1	43.1	44.6	47.4	43.0	42.0	40.4	47.3	.871	.871	.872	.868
AR	44.7	41.9	43.8	47.3	41.5	39.7	38.3	47.1	.863	.860	.868	.849
WK	22.8	21.4	23.0	23.6	22.0	21.0	20.9	23.3	.822	.817	.829	.799
PC	98.4	88.7	91.3	111.2	90.8	83.6	79.7	110.8	.749	.745	.754	.729
AI	60.0	58.5	57.8	66.7	56.7	54.8	51.8	66.1	.863	.859	.866	.851
SI	57.3	57.2	56.4	62.9	52.3	50.9	48.1	60.7	.872	.870	.875	.862
MK	67.2	65.0	63.4	78.4	63.3	63.6	58.0	76.8	.904	.902	.906	.896
MC	73.8	67.1	68.0	83.8	68.9	63.1	61.3	82.8	.824	.821	.828	.810
EI	73.0	73.4	71.0	83.0	66.5	65.2	60.1	79.9	.835	.833	.839	.823

TABLE 4

PROPERTIES OF ESTIMATORS IN θ METRIC WITH *N*(0,1) PRIOR DISTRIBUTION, MODEL ASSUMPTIONS VIOLATED BECAUSE OF EFFECT OF MEDIUM OF ADMINISTRATION[a]

	Mean squared error of estimation			Variance of measurement error			Variance of true scores			Squared correlation with θ			Reliability		
Sub-test	Mean	Mode	Owen	Mean	Mode	Owen	Mean	Mode	Owen	Mean	Mode	Owen	Mean	Mode	Owen
GS	6.8	6.6	7.1	6.4	6.3	6.7	31.6	30.8	32.7	.831	.830	.829	.832	.831	.830
AR	6.4	6.0	6.8	6.1	5.6	6.4	34.2	31.6	36.0	.848	.848	.846	.849	.849	.849
WK	5.2	5.0	5.4	4.8	4.6	5.1	26.2	25.2	27.4	.845	.845	.842	.846	.846	.844
PC	11.8	10.7	12.7	11.5	10.1	12.2	29.4	25.5	31.3	.718	.716	.717	.718	.716	.720
AI	7.6	7.4	8.1	7.2	6.9	7.3	48.9	46.6	49.3	.868	.866	.861	.872	.870	.870
SI	10.2	10.3	10.4	9.3	9.0	9.1	59.5	57.1	58.0	.865	.864	.861	.865	.864	.864
MK	6.1	5.8	6.5	5.7	5.5	6.0	36.2	34.2	37.4	.863	.861	.859	.864	.861	.861
MC	10.6	9.4	10.8	9.9	9.0	10.0	33.4	29.3	33.1	.769	.764	.768	.771	.766	.769
EI	11.6	11.6	12.0	10.3	10.2	10.2	44.8	44.2	43.8	.811	.809	.804	.813	.812	.811

[a] Mean squared errors and variances have been multiplied by 100.

TABLE 5

PROPERTIES OF ESTIMATORS IN PERCENT-CORRECT METRIC WITH *N*(0,1) PRIOR DISTRIBUTION, MODEL ASSUMPTIONS VIOLATED BECAUSE OF EFFECT OF MEDIUM OF ADMINISTRATION

	Mean squared error of estimate				Variance of measurement error				Reliability			
Subtest	P(mean)	P(mode)	Mean(P)	Mode(P)	P(mean)	P(mode)	Mean(P)	Mode(P)	P(mean)	P(mode)	Mean(P)	Mode(P)
GS	74.5	71.7	72.8	78.6	69.0	67.4	64.2	77.8	.794	.793	.798	.785
AR	59.8	55.7	58.3	62.6	56.5	52.7	52.5	61.9	.814	.813	.818	.802
WK	24.3	22.5	24.7	26.1	22.7	19.8	22.1	22.9	.778	.783	.784	.757
PC	112.2	105.4	105.7	127.5	102.6	94.0	90.3	123.0	.693	.689	.698	.673
AI	75.8	72.0	72.0	84.3	69.4	67.2	62.9	82.9	.840	.836	.844	.824
SI	77.3	79.2	74.4	89.3	71.2	68.8	65.3	82.1	.825	.822	.829	.812
MK	113.0	110.0	106.8	127.7	103.7	104.3	94.8	124.8	.843	.839	.846	.831
MC	91.3	84.4	83.9	106.3	86.9	81.0	77.6	105.1	.781	.775	.783	.768
EI	98.7	97.3	96.1	108.6	86.2	84.8	78.0	104.4	.784	.782	.788	.770

reduces reliabilities. Changes in item parameters increase error variance and, in general, also reduce true-score variance.

The loss of reliability is presented in table 6 for the modal estimate, which has been approved for use in CAT-ASVAB ([3 (item 6)]. Following the CAT-ASVAB plans, the estimator in percent-correct metric is *P(mode).* The first line in table 6 quantifies the size of the medium effect, in terms of mean average absolute difference (AAD) between CAT and PP item characteristic curves [9 (table 2)].

TABLE 6

SIZE OF MEDIUM EFFECT AND CONSEQUENT DECREASES IN RELIABILITY IN θ AND PERCENT-CORRECT METRICS

	Subtest								
	GS	AR	WK	PC	AI	SI	MK	MC	EI
Mean AAD	.050	.047	.048	.051	.061	.079	.064	.089	.071
θ metric	.053	.048	.047	.050	.013	.035	.048	.061	.049
% metric	.078	.048	.034	.056	.023	.048	.063	.046	.052

Given that the occupational subtests AI, SI, MC, and EI are more sensitive to the medium effect than the academic subtests, it is surprising that the degradation in reliability is about the same for both types.

CONCLUSIONS

It is clear that the smaller MSE of the posterior mean does not translate into noticeably superior reliability. This supports the decision to use the modal estimate in CAT-ASVAB [3]. In fact, except for its dependence on item order, even Owen's estimate would be satisfactory. Results obtained recently by Sympson [3] support the use of Owen's approximation for item selection.

The other major conclusion is that CAT reliability drops appreciably if the medium of administration changes item parameters on a scale found in the experimental CAT-ASVAB data [8, 9]. Therefore simulations without a medium effect overestimate

CAT reliability. More realistic simulations, allowing for changes in item parameters from PP to CAT, cannot be performed until enough CAT data are in hand to permit estimation of item parameters. In the meantime, all one can do is refrain from making strong claims about the reliability of the CAT version.

REFERENCES

[1] Navy Personnel Research and Development Laboratory, TR 84-33, *Microcomputer Network for Computerized Adaptive Testing (CAT)*, by Baldwin Quan, Thomas A. Park, Gary Sandahl, and John H. Wolfe, Mar 1984

[2] Roger J. Owen. "A Bayesian Sequential Procedure for Quantal Response in the Context of Adaptive Mental Testing." *Journal of the American Statistical Association* (June 1975): 351-356.

[3] Defense Manpower Data Center, *Minutes of January 1987 Meeting*, by Bruce Bloxom, 15 Feb 1987

[4] CNA Research Memorandum 86-46, *On Inaccuracies in Owen's Approximation for the Bayesian Ability Estimate*, by D. R. Divgi, Feb 1986

[5] Naval Postgraduate School, *Minutes of June 1986 Meeting*, by Bruce Bloxom, 1 Aug 1986

[6] R. Darrell Bock and Robert J. Mislevy, "Adaptive EAP Estimation of Ability in a Microcomputer Environment," *Applied Psychological Measurement* (Fall 1982): 431-444

[7] J. B. Sympson. *Bayesian Estimation of True Scores and Observed Scores on a Criterion Test*, paper presented at the Annual Meeting of the Psychometric Society, Jun 1985

[8] CNA Research Memorandum 86-24, *Effect of the Medium of Administration on ASVAB Item Response Curves*, by D. R. Divgi and Peter H. Stoloff, Apr 1986

[9] CNA Research Memorandum 86-189, *Sensitivity of CAT-ASVAB Scores to Changes in Item Response Curves with the Medium of Administration*, by D. R. Divgi, Aug 1986